Navigating Safe Campus Operations during Epidemics with Reinforcement Learning

Elizabeth Akinyi Ondula
ondula@usc.edu
Viterbi School of Engineering
University of Southern California

Bhaskar Krishnamachari
bkrishna@usc.edu
Viterbi School of Engineering
University of Southern California

Abstract

Epidemic modeling, which includes both deterministic and stochastic methods, has been central to understanding infectious disease dynamics and guiding public health decisions. While a significant portion of machine learning research in this domain focuses on predictions and trends of the disease, this study takes a prescriptive approach. This work introduces SafeCampus [1], a tool that simulates infection spread and facilitates the exploration of various RL algorithms in response to epidemic challenges. The focus is in using reinforcement learning (RL) to develop occupancy strategies that could balance minimizing infections with maximizing in-person interactions in educational settings. SafeCampus incorporates a custom RL environment, leveraging a stochastic epidemic model, to realistically represent university campus dynamics during epidemics. We evaluate a Q-learning algorithm in this context for a discretized state space to yield a sensible policy matrix, which prescribes decisions about the level of occupancy suitable for different epidemiological phases.

1 Introduction

Traditional epidemic responses often struggle to adapt to uncertainties, leading to compromised effectiveness and operational challenges Barnett et al. (2023). The COVID-19 pandemic in 2020 highlighted this issue, causing a global shutdown of education systems Bank (2020). These static strategies require reevaluation, as educational environments demand innovative approaches that adapt to public health threats while supporting education. This research focuses on determining optimal classroom occupancy levels during an epidemic. Localized strategies offer the flexibility to adapt to unique campus dynamics and rapidly changing epidemic conditions.

To address these challenges, we propose a reinforcement learning environment using a discrete-time approximate SI model to simulate campus dynamics during an epidemic. The Q-learning agent interacts with the environment, making decisions about student attendance percentages and receiving feedback as rewards. The central problem aims to ensure safe campus operations while balancing in-person attendance for educational benefits. The agent must learn this policy through Q-learning.

This work makes the following contributions:

- First, we present SafeCampus, a tool designed to incorporate a variety of stochastic epidemic models to simulate a range of infection scenarios. This could allow for studying different aspects of epidemic spread and control for indoor spaces.
- Second, we introduce the use of a discrete-time approximate SI model within the reinforcement learning environment, providing a computationally efficient and flexible approach to modeling the dynamics of indoor environments during an epidemic.

[1] https://github.com/ANRGUSC/SafeCampus

- Third, we identify policies that can help find a balance between the delicate balance between e.g educational benefit of in-person attendance and health safety.

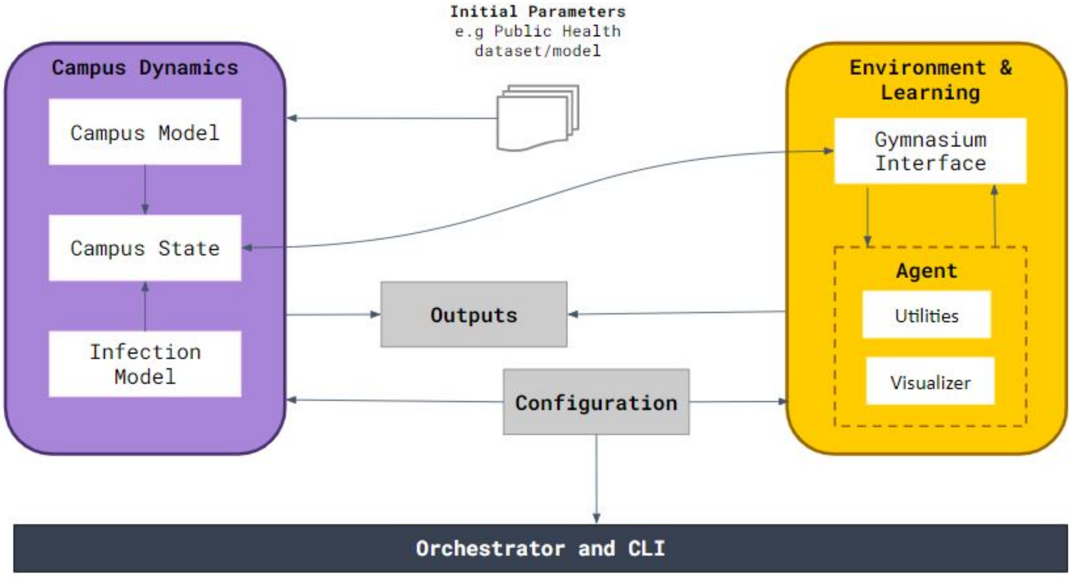


Figure 1: *A schematic of the system, integrating campus dynamics, agent, and orchestration components.*

2 Related Works

Reinforcement learning (RL) has been widely applied in healthcare, economics, and mobility for epidemic control. Studies by Arango and Pelov (2020), Ohi et al. (2020), Feng et al. (2022), Probert et al. (2019), Bushaj et al. (2023), Kompella et al. (2020), and Uddin et al. (2020) use algorithms like Deep Q-Learning and Proximal Policy Optimization (PPO) with compartmental models (e.g., SEIR, SIR, SIHR) and granular models like agent-based and meta-population models. These enable RL agents to learn in realistic epidemic environments, optimizing policies for mobility restrictions, lockdowns, testing, sanitization, social distancing, ventilation control, and vaccine distribution to balance infection control with socioeconomic impacts.

Our work focuses on the unique dynamics of university campuses, differing from general public health applications. Existing studies often implement broader strategies, whereas we aim to develop dynamic strategies that adjust student attendance to balance campus operations with health safety. We present SafeCampus, incorporating various stochastic epidemic models to simulate infection scenarios and study epidemic control within educational contexts.

We integrate stochastic epidemic models with RL to derive policies that maximize social interactions while limiting disease spread. Unlike previous studies focused on lockdowns, we explore flexible, campus-level policies using a discrete-time approximate SI model within the RL environment for efficient and flexible modeling of indoor epidemic dynamics.

Research at the intersection of epidemics and education, particularly during COVID-19, has explored strategies to mitigate disease spread. Studies by Fukumoto et al. (2021) and Wu et al. (2022) questioned school closures' effectiveness, noting children's lower susceptibility. Best et al. (2021) and Kaiser et al. (2020) found that smaller class sizes and cohorting limit outbreaks. Haelermans et al. (2022) highlighted closures' adverse impact on disadvantaged groups' learning progress. Historically,

Table 1: Related Works Comparing RL Methods and Optimization Goals

	Optimization Goals		
	Healthcare System Efficiency	**Economic Optimization**	**Policy Development**
RL Method			
(Q/SARSA) Learning/Actor Critic	Arango and Pelov (2020), Deng et al. (2021), Khatami and Gopalappa (2022), Kompella et al. (2020)	Guo et al. (2022), Khadilkar et al. (2020), Ohi et al. (2020)	Probert et al. (2019), Wang et al. (2023)
Proximal Policy Optimization (PPO)		Feng et al. (2022)	Feng et al. (2023), Hosseinloo et al. (2022), Libin et al. (2021), Mai et al. (2023)
Hierarchical RL		Uddin et al. (2020)	Hao et al. (2021), Du et al. (2023)

as noted by Spielman and Sunavala-Dossabhoy (2021), epidemics have accelerated digital learning. Current research, including Endo et al. (2022) and Oikawa et al. (2022), emphasizes a multifaceted approach to limit spread within schools. Our research focuses on balancing educational benefits of in-person attendance with health safety, addressing the challenges of maintaining campus operations during an epidemic.

3 Modeling

3.1 Problem Definition

Consider a classroom scenario with N students attending sessions over W weeks. During an ongoing epidemic, students face the risk of infection both off-campus and on-campus. Off-campus infections are considered to be an exogenous random process, where each student has an independent and identical probability c_w of being infected off-campus during week w, termed the *community risk*. On-campus infections result from infected students spreading a virus to other students.

3.2 Approximate SI(Susceptible-Infectious) Model

Conventional compartmental models, such as the SI model, describe epidemic dynamics using coupled ordinary differential equations that capture the instantaneous rates of change in the susceptible and infected populations. In contrast, our approximate SI model simplifies this approach by using a discrete-time framework, which better suits the discrete nature of policy decisions, such as adjusting student attendance percentages. We use a discrete-time model as it allows for simulating the impact of interventions at specific time points, aligning with the discrete nature of policy decisions and enabling the evaluation of different intervention scenarios. This model employs recursive equations that update the infected populations at each time step based on the previous state and proportional relationships.

Let N represent the total number of students allowed in a campus classroom, and I_{t-1} denote the number of infected students from the previous time step. The community risk of infection is represented by c_r. The constants ϕ and β represent the infection rates inside the campus and due to community interactions, respectively. The total number of students considered in the model is N. The number of new infections I_t at time t is then estimated using the following relation:

$$I_t = \min\left(\left[\phi \cdot I_{t-1} \cdot A_t + \beta \cdot c_r \cdot A_t^2\right], A_t\right) \tag{1}$$

where A = the number of allowed (susceptible) population. The term $\phi \cdot I_{t-1}$ represents new infections inside a classroom, while $\beta \cdot c_r \cdot S_t^2$ represents new infections due to interactions outside the classroom. $c_r t$ represents community risk.

Parameter	Description	Value
I	**Number of infected population**	**Max** $= N$
A_t	**Number of population allowed in the classroom at time** t	**0, 50, or 100**
c_r	**Community infection risk factor**	**range (0,1)**
ϕ	**Constant representing the indoor transmission risk**	**0.005**
β	**Representing the scale effect from the community**	**0.01**
N	**Total number of students considered in the model**	**100**

Table 2: Description of Parameters

3.3 Role of ϕ and β

The parameter ϕ represents the indoor transmission risk. Some factors that could influence this are mask mandates or social distancing rules, which directly impact the infection rate within the classroom. Effective implementation of these measures can reduce the value of ϕ, thereby lowering the transmission risk in indoor environments.

On the other hand, β represents the scale effect from the community. This parameter includes factors such as public health guidelines and the overall level of community transmission. A high value of β suggests that despite stringent indoor policies, the external risk remains significant, potentially due to high community transmission rates or insufficient adherence to public health measures outside the classroom environment.

3.4 Model Dynamics

We conducted a parameter sweep analysis to investigate the impact of varying levels of allowed interactions (A_t) and community risk ($c_r t$) on peak infection numbers within a campus population. The model simulates the infection dynamics over a series of discrete time steps, updating the number of infected individuals using the infection model equation. Our findings indicate that increasing A_w interactions or c_w significantly contributes to higher peak infection numbers.

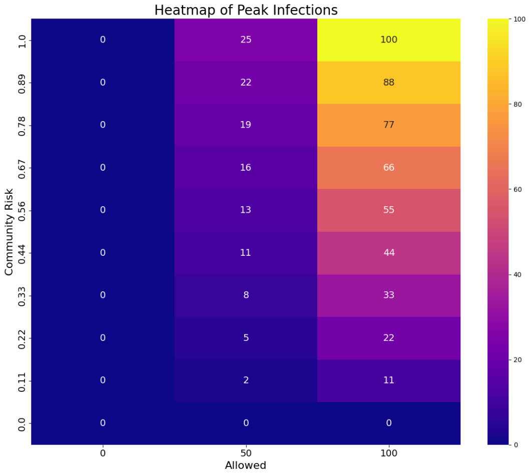


Figure 2: Heatmap of Peak Infections

4 The Reinforcement Learning Problem

We formulate the problem of finding operational strategies during an ongoing epidemic as a model-free RL problem where we are interested in developing a policies that would take a given set of observations about the infection process and make a decision on how many students to allow in the classroom at the beginning of every week w. The RL problem is thus formulated as follows:

- **State space:** We use as the state observation a tuple consisting of the community risk and the current number of (expected) infected students, i.e. $(c_w,\ E[I_w])$. For simplicity and efficiency, we are currently discretizing the observed state space into a set of discrete levels for both c_w, $E[I_w]$. A range of 1-10 is used and this could be easily modified to accommodate a more fine-grained discretization at the expense of greater storage and computational complexity for the reinforcement learning.
- **Action space:** The output of the policy A_w is the number of students allowed to participate in the class. Again, for ease of implementation, we discretize the action into L levels (e.g., if $L = 3$, the possible actions may be to allow 0, $0.5N$ or N students in a given week).
- **State-transition model:** Within our environment, this is governed by the dynamics of the approximate SI model described earlier which simulates the spread of infection in a classroom setting.
- **Reward:** The function is designed to capture the trade-off between the benefit of in-person interactions and the cost associated with infection risk. It considers community risk, the number of allowed students, and the current number of infected students. It is defined as:

$$\text{Reward} = \text{int}\left(\alpha_r \times A_{\text{allowed}} - ((1-\alpha) \times I_{\text{current}})\right) \tag{2}$$

  where: I_{current} is the number of infected students at week, A_{allowed} is the number of students allowed, and α is a weighting factor that balances the priority between maximizing student attendance A_{allowed} and minimizing the number of infections).

4.1 Reward Assumption

α is a parameter that is to be determined by a human operator such as a a campus administrator in this context. α may depend on factors such as severity of a disease or class type such as a lab or lecture. A higher α places greater emphasis on increasing attendance, while a lower α gives more weight to reducing infections, thus allowing the Q-learning algorithm to prioritize between educational benefits and health risks according to the chosen value of α.

4.2 Temporal Difference Learning

Temporal Difference (TD) Learning is a central concept in reinforcement learning, combining ideas from both Monte Carlo methods and dynamic programming. The value of a state (or state-action pair) is updated using the difference between the predicted value of the current state and the value of the next state, adjusted by the reward received in transitioning between these states. This difference is known as the TD error.

We apply Q-learning Watkins and Dayan (1992), an off-policy TD control algorithm, in this context since it does not require a predefined model of the environment's dynamics, making it suitable for situations where the exact mechanisms of infection spread are complex or not fully understood. Q-learning learns from interactions with the environment, gaining knowledge directly through trial and error.

The TD update rule for Q-learning is given by the Bellman equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right] \tag{3}$$

where:

- $Q(s, a)$ is the current estimate of the Q-value for the state-action pair (s, a),
- r is the reward received after executing action a from state s,
- γ is the discount factor,
- α is the learning rate,
- s' is the next state,
- $\max_{a'} Q(s', a')$ is the maximum estimated Q-value for the next state s' across all possible actions a'.

Q-learning evaluates the target policy while following a behavior policy, thus it is an off-policy algorithm. This means that the policy used to generate behavior (behavior policy) is different from the policy that is being improved and evaluated (target policy). It estimates the Q-values for the target (optimal) policy using trajectories from a behavior policy. It converges to the optimal policy in finite state-action spaces, assuming all state-action pairs are explored. However, in continuous state spaces, the curse of dimensionality limits its efficiency and convergence. In this context, we are specifically using tabular Q-learning.

Algorithm 1 Q-Learning

```
Initialize empty Q-table
for episode ← 1 to max_episodes do
    state ← reset environment
    terminated ← False
    while not terminated do
        Choose an action based on current policy
        Execute action and observe reward, next state
        Update Q-table according to:
        $Q(s, a) \leftarrow Q(s, a) + \alpha\, [r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
        Update state to next state
    end while
    Update policy based on learned Q-values
    Decay exploration and learning rates as needed
end for
```

5 Tool Overview

SafeCampus is engineered to simulate a range of scenarios, enabling the evaluation of various policy decisions concerning infection control and in-person interactions. Figure 1 shows the Key components that includes the Campus Dynamics modules, which are essential in defining and managing the evolving state of the campus environment. This includes the `campus model`, `campus state`, and `infection model`, each responsible for initializing the simulation, managing dynamic system states, and simulating infection spread using various epidemiological models. The Environment and Learning component integrates these behaviors to optimize strategies through learned experiences in an RL framework, utilizing tools like the `gymnasium interface` for agent interactions and an `agent package` for implementing various RL algorithms, primarily focusing on Q-learning. Supplementar modules, such as the `configuration`, `outputs`, and `orchestration` modules, provide essential support in system configuration, data management, and overall system control, respectively. Furthermore, the `command-line interface (CLI)` facilitates user interaction, allowing for precise control over operational modes and parameters, thus driving the system's training, evaluation, and optimization processes.

6 Experiments

In our study, we investigate the following research questions:

1. How does the reward weight parameter α_r influence the Q-learning agent's ability to develop optimal occupancy policies in response to varying infection counts and community risk patterns?
2. Can Q-learning generate a sensible policy matrix that prescribes specific occupancy decisions effectively under dynamic and uncertain infection scenarios?
3. How well does the Q-learning agent, trained in a simulation environment with simulated community risk values, generalize and adapt its decision-making strategy when applied to real-world risk level data?

Hypotheses: We hypothesize that Q-learning can effectively generate a sensible policy matrix that prescribes specific occupancy decisions based on infection counts and community risk patterns. We also posit that the reward weight parameter α_r will play a crucial role in shaping the Q-learning agent outcomes and the precision of the policy matrix. By varying the α_r value, our objective is to explore how the algorithm prioritizes educational benefits by *allowing more students* and infection risk minimization, thereby calibrating the matrix to align with varying epidemic scenarios.

Training with different α_r values: To explore how the reward weight parameter α_r influences the Q-learning agent's ability to develop optimal occupancy policies, we conducted experiments by varying α_r values. This approach allowed us to systematically analyze the trade-off between in-person learning and the risk of infection due to increased physical interactions.

7 Evaluation

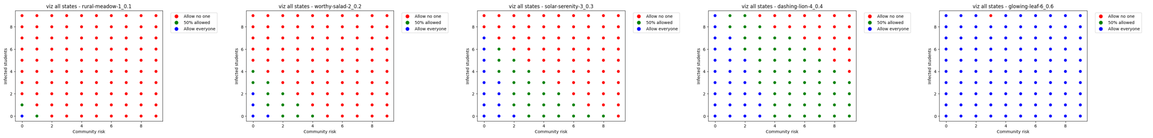
Metrics: To assess the effectiveness of the Q-learning agent in developing optimal occupancy policies, we evaluated the agent based on the following metrics:

- **Policy Accuracy**: The accuracy of the generated policy matrix in prescribing occupancy decisions that align with infection counts and community risk patterns.
- **Adaptability**: The ability of the Q-learning agent to generalize and adapt its decision-making strategy when applied to real-world risk level data.
- **Return (Reward) Analysis**: The moving average of the expected return (reward) over episodes to determine the stability and convergence of the learning process under different α_r values.

Policy Visualization: In Figure 3, each matrix depicts the algorithm's occupancy recommendations across different states of community risk and infection counts. The policy gradient shifts from conservative (red dots: Allow no one) to permissive (blue dots: Allow everyone) as the α_r value increases, indicating a higher emphasis on educational benefits. Green dots represent a balanced occupancy decision (50% allowed). The progression from (a) to (i) captures the algorithm's adaptive responses to the campus dynamics, showcasing the delicate balance between ensuring educational benefits and managing infection risks.

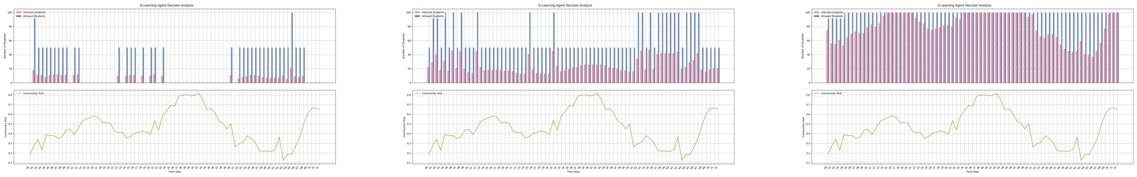
7.1 Sim-to-Real-World Data

We further evaluated the Q-learning agent's ability to generalize and adapt its decision-making strategy using real-world COVID-19 risk score data Kiamari et al. (2020). We resampled the data to a weekly frequency to align with the academic semester simulation timeframe. Then calculated the mean of these weekly normalized risk levels across all regions to generate a single aggregated risk level for each week. At a low α_r (0.2), the agent adopts a conservative strategy (Figure 4(a)), allowing fewer students to attend in-person classes, which results in lower infection rates but limits educational interactions, effectively prioritizing safety under high-risk conditions. With a medium α_r (0.4), the agent achieves a balance between infection risk and occupancy (Figure 4(b)), resulting in moderate infection rates and demonstrating adaptability to dynamic risk levels. At a high α_r (0.6), the agent allows more students to attend in-person classes (Figure 4(c)), accepting higher infection risks to maximize educational interactions, reflecting its flexibility in less conservative policies when community risk is perceived to be lower.



((a)) $\alpha_r = 0.1$ ((b)) $\alpha_r = 0.2$ ((c)) $\alpha_r = 0.3$ ((d)) $\alpha_r = 0.4$ ((e)) $\alpha_r = 0.5$

Figure 3: Policy matrices for different α_r values. Red dots: Allow no one, Green dots: 50% allowed, Blue dots: Allow everyone.



((a)) $\alpha_r = 0.2$ ((b)) $\alpha_r = 0.4$ ((c)) $\alpha_r = 0.6$

Figure 4: The blue bars represent the agent action of allowing, red bars represent the infected. The bottom graph is the aggregated community risk values from the COVID-19 risk scores

8 Conclusion

Our findings suggest that Q-learning can effectively navigate the trade-offs involved in epidemic management within educational settings. By systematically adjusting operational policies based on data and outcomes, a Q-learning agent demonstrates a capacity for nuanced decision-making, ensuring safety while minimizing disruption to educational processes.

Limitations and Future Work: Despite the promising results, there are limitations to our approach. The tabular Q-learning method may struggle with larger, continuous state spaces due to the curse of dimensionality. Future work could explore function approximation techniques, such as Deep Q-Networks (DQN), to handle more complex environments. Additionally, integrating domain expertise into the reward function design could enhance the alignment of the learned policies with real-world operational goals.

References

Mauricio Arango and Lyudmil Pelov. Covid-19 pandemic cyclic lockdown optimization using reinforcement learning. *arXiv preprint arXiv:2009.04647*, 2020.

World Bank. *The COVID-19 pandemic: Shocks to education and policy responses*. World Bank, 2020.

Michael Barnett, Greg Buchak, and Constantine Yannelis. Epidemic responses under uncertainty. *Proceedings of the National Academy of Sciences*, 120(2):e2208111120, 2023.

Alex Best, Prerna Singh, Charlotte Ward, Caterina Vitale, Megan Oliver, Laminu Idris, and Alison Poulston. The impact of varying class sizes on epidemic spread in a university population. *Royal Society Open Science*, 8(6):210712, 2021.

Sabah Bushaj, Xuecheng Yin, Arjeta Beqiri, Donald Andrews, and İ Esra Büyüktahtakın. A simulation-deep reinforcement learning (sirl) approach for epidemic control optimization. *Annals of Operations Research*, 328(1):245–277, 2023.

Wei Deng, Guoyuan Qi, and Xinchen Yu. Optimal control strategy for covid-19 concerning both life and economy based on deep reinforcement learning. *Chinese Physics B*, 30(12):120203, 2021.

Xinqi Du, Hechang Chen, Bo Yang, Cheng Long, and Songwei Zhao. Hrl4ec: Hierarchical reinforcement learning for multi-mode epidemic control. *Information Sciences*, 640:119065, 2023.

Akira Endo, CMMID COVID-19 Working Group, Mitsuo Uchida, Yang Liu, Katherine E Atkins, Adam J Kucharski, and Sebastian Funk. Simulating respiratory disease transmission within and between classrooms to assess pandemic management strategies at schools. *Proceedings of the National Academy of Sciences*, 119(37):e2203019119, 2022.

Tao Feng, Tong Xia, Xiaochen Fan, Huandong Wang, Zefang Zong, and Yong Li. Precise mobility intervention for epidemic control using unobservable information via deep reinforcement learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2882–2892, 2022.

Tao Feng, Sirui Song, Tong Xia, and Yong Li. Contact tracing and epidemic intervention via deep reinforcement learning. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–24, 2023.

Kentaro Fukumoto, Charles T McClean, and Kuninori Nakagawa. Shut down schools, knock down the virus? no causal effect of school closures on the spread of covid-19. *medRxiv*, pages 2021–04, 2021.

Xudong Guo, Peiyu Chen, Shihao Liang, Zengtao Jiao, Linfeng Li, Jun Yan, Yadong Huang, Yi Liu, and Wenhui Fan. Pacar: Covid-19 pandemic control decision making via large-scale agent-based modeling and deep reinforcement learning. *Medical Decision Making*, 42(8):1064–1077, 2022.

Carla Haelermans, Madelon Jacobs, Rolf van der Velden, Lynn van Vugt, and Sanne van Wetten. Inequality in the effects of primary school closures due to the covid-19 pandemic: Evidence from the netherlands. In *AEA Papers and Proceedings*, volume 112, pages 303–307. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2022.

Qianyue Hao, Fengli Xu, Lin Chen, Pan Hui, and Yong Li. Hierarchical reinforcement learning for scarce medical resource allocation with imperfect information. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2955–2963, 2021.

Ashkan Haji Hosseinloo, Saleh Nabi, Anette Hosoi, and Munther A Dahleh. Data-driven control of covid-19 in buildings: a reinforcement-learning approach. *arXiv preprint arXiv:2212.13559*, 2022.

Anna Kaiser, David Kretschmer, and Lars Leszczensky. Social network-based strategies for classroom size reduction can help limit outbreaks of sars-cov-2 in high schools. a simulation study in classrooms of four european countries. *medRxiv*, pages 2020–11, 2020.

Harshad Khadilkar, Tanuja Ganu, and Deva P Seetharam. Optimising lockdown policies for epidemic control using reinforcement learning: An ai-driven control approach compatible with existing disease and network models. *Transactions of the Indian National Academy of Engineering*, 5(2): 129–132, 2020.

Seyedeh Nazanin Khatami and Chaitra Gopalappa. Deep reinforcement learning framework for controlling infectious disease outbreaks in the context of multi-jurisdictions. *medRxiv*, pages 2022–10, 2022.

Mehrdad Kiamari, Gowri Ramachandran, Quynh Nguyen, Eva Pereira, Jeanne Holm, and Bhaskar Krishnamachari. Covid-19 risk estimation using a time-varying sir-model. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19*, pages 36–42, 2020.

Varun Kompella, Roberto Capobianco, Stacy Jong, Jonathan Browne, Spencer Fox, Lauren Meyers, Peter Wurman, and Peter Stone. Reinforcement learning for optimization of covid-19 mitigation policies. *arXiv preprint arXiv:2010.10560*, 2020.

Pieter JK Libin, Arno Moonens, Timothy Verstraeten, Fabian Perez-Sanjines, Niel Hens, Philippe Lemey, and Ann Nowé. Deep reinforcement learning for large-scale epidemic control. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 155–170. Springer, 2021.

Anh Mai, Nikunj Gupta, Azza Abouzied, and Dennis Shasha. Planning multiple epidemic interventions with reinforcement learning. *arXiv preprint arXiv:2301.12802*, 2023.

Abu Quwsar Ohi, MF Mridha, Muhammad Mostafa Monowar, and Md Abdul Hamid. Exploring optimal control of epidemic spread using reinforcement learning. *Scientific reports*, 10(1):22106, 2020.

Masato Oikawa, Ryuichi Tanaka, Shun-ichiro Bessho, and Haruko Noguchi. Do class size reductions protect students from infectious diseases? lessons for covid-19 policy from a flu epidemic in the tokyo metropolitan area. *American Journal of Health Economics*, 8(4):449–476, 2022.

William JM Probert, Sandya Lakkur, Christopher J Fonnesbeck, Katriona Shea, Michael C Runge, Michael J Tildesley, and Matthew J Ferrari. Context matters: using reinforcement learning to develop human-readable, state-dependent outbreak response policies. *Philosophical Transactions of the Royal Society B*, 374(1776):20180277, 2019.

Andrew I Spielman and Gulshan Sunavala-Dossabhoy. Pandemics and education: A historical review. *Journal of dental education*, 85(6):741–746, 2021.

M Irfan Uddin, Syed Atif Ali Shah, Mahmoud Ahmad Al-Khasawneh, Ala Abdulsalam Alarood, and Eesa Alsolami. Optimal policy learning for covid-19 prevention using reinforcement learning. *Journal of Information Science*, page 0165551520959798, 2020.

Xintong Wang, Gary Qiurui Ma, Alon Eden, Clara Li, Alexander Trott, Stephan Zheng, and David Parkes. Platform behavior under market shocks: A simulation framework and reinforcement-learning based study. In *Proceedings of the ACM Web Conference 2023*, pages 3592–3602, 2023.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

Joseph T Wu, Shujiang Mei, Sihui Luo, Kathy Leung, Di Liu, Qiuying Lv, Jian Liu, Yuan Li, Kiesha Prem, Mark Jit, et al. A global assessment of the impact of school closure in reducing covid-19 spread. *Philosophical Transactions of the Royal Society A*, 380(2214):20210124, 2022.